Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	937	web adj crawl$	US-PGPUB; USPAT; IBM_TDB	OR	ON	2005/12/19 13:40
L2	13	1 with statistic$	US-PGPUB; USPAT; IBM_TDB	OR	ON	2005/12/19 15:07
L3	44	1 same statistic$	US-PGPUB; USPAT; IBM_TDB	OR	ON	2005/12/19 15:07
L4	31	3 not 2	US-PGPUB; USPAT; IBM_TDB	OR	ON	2005/12/19 15:08
L5	5	("6418433").URPN.	USPAT	OR	ON	2005/12/19 15:12
L6	1	("6615259").URPN.	USPAT	OR	ON	2005/12/19 16:46
L7	1	("6671723").URPN.	USPAT	OR	ON	2005/12/19 16:47
L8	64	(user adj defin$) with predicate	USPAT	OR	ON	2005/12/19 18:18
L9	2	((user adj defin$) with predicate) and (statistic$ near information)	USPAT	OR	ON	2005/12/19 18:18
L10	1	"6343288".URPN.	USPAT	OR	ON	2005/12/19 18:18
L11	2	((user adj defin$) with predicate) and (web adj page)	USPAT	OR	ON	2005/12/19 18:18
L12	2	((user adj defin$) with predicate) and ((web adj page) html)	USPAT	OR	ON	2005/12/19 18:18
L13	2	((user adj defin$) with predicate) and ((webpage) html)	USPAT	OR	ON	2005/12/19 18:18
L14	2714	predicate	USPAT	OR	ON	2005/12/19 18:18
L15	111	predicate and (statistic$ near information)	USPAT	OR	ON	2005/12/19 18:18
L16	40	(predicate and (statistic$ near information)) and (web adj page)	USPAT	OR	ON	2005/12/19 18:18
L17	36485	((uniform adj resource adj locator) url) token	USPAT	OR	ON	2005/12/19 18:18
L18	114	((uniform adj resource adj locator) url) with token	USPAT	OR	ON	2005/12/19 18:18
L19	0	((predicate and (statistic$ near information)) and (web adj page)) and (((uniform adj resource adj locator) url) with token)	USPAT	OR	ON	2005/12/19 18:18
L20	0	(predicate and (statistic$ near information)) and (((uniform adj resource adj locator) url) with token)	USPAT	OR	ON	2005/12/19 18:18

L21	7	predicate and (((uniform adj resource adj locator) url) with token)	USPAT	OR	ON	2005/12/19 18:18
L22	39	((predicate and (statistic$ near information)) and (web adj page)) and (retriev$ with (document page file))	USPAT	OR	ON	2005/12/19 18:18
L23	38	((predicate and (statistic$ near information)) and (web adj page)) and (retriev$ with (document))	USPAT	OR	ON	2005/12/19 18:18
L24	3284	(707/3).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L25	1314	(707/6).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L26	578	(715/530).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L27	1121	(715/513).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L28	716	(715/501.1).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L29	164	(((715/530).CCLS.) ((715/513). CCLS.)) and (((707/3).CCLS.) ((707/6).CCLS.))	USPAT	OR	ON	2005/12/19 18:18
L30	31	(((715/530).CCLS.)) and (((707/3).CCLS.) ((707/6).CCLS.))	USPAT	OR	ON	2005/12/19 18:18
L31	138	(((715/513).CCLS.)) and (((707/3).CCLS.) ((707/6).CCLS.))	USPAT	OR	ON	2005/12/19 18:18
L32	413	(user adj defin$) with (query)	USPAT	OR	ON	2005/12/19 18:18
L33	4672	statistic$ near information	USPAT	OR	ON	2005/12/19 18:18
L34	23	((user adj defin$) with (query)) and (statistic$ near information)	USPAT	OR	ON	2005/12/19 18:18
L35	80	((query and (statistic$ near information)) and (web adj page)) and (retriev$ with (document))	USPAT	OR	ON	2005/12/19 18:18
L36	10	(((user adj defin$) with (query)) and (statistic$ near information)) and ((web adj page) html webpage)	USPAT	OR	ON	2005/12/19 18:18
L37	5144	search near3 (query term phrase)	USPAT	OR	ON	2005/12/19 18:18
L38	613	(search near3 (query term phrase)) same (retriev$ with (document file webpage (web adj page) page))	USPAT	OR	ON	2005/12/19 18:18
L39	233	((search near3 (query term phrase)) same (retriev$ with (document file webpage (web adj page) page))) and (statistic$)	USPAT	OR	ON	2005/12/19 18:18

L40	782	(URL (uniform adj resource adj locator)) near3 (token$ string$)	USPAT	OR	ON	2005/12/19 18:18
L41	9	(((search near3 (query term phrase)) same (retriev$ with (document file webpage (web adj page) page))) and (statistic$)) and ((URL (uniform adj resource adj locator)) near3 (token$ string$))	USPAT	OR	ON	2005/12/19 18:18
L42	3006	search near (query term phrase)	USPAT	OR	ON	2005/12/19 18:18
L43	59	(search near (query term phrase)) same (retriev$ with (webpage (web adj page) HTML))	USPAT	OR	ON	2005/12/19 18:18
L44	2	((search near (query term phrase)) same (retriev$ with (webpage (web adj page) HTML))) and (statistic$ near information)	USPAT	OR	ON	2005/12/19 18:18
L45	5	((search near (query term phrase)) same (retriev$ with (webpage (web adj page) HTML))) and (statistic$ with information)	USPAT	OR	ON	2005/12/19 18:18
L46	1644	(707/4).CCLS.	USPAT; USOCR	OR	OFF	2005/12/19 18:18
L47	9	("5619709" | "5692176" | "5717914" | "5737734" | "5926812" | "6212532" | "6272495" | "6353823" | "6363377").PN.	USPAT	OR	ON	2005/12/19 18:18
L48	0	"6633868".URPN.	USPAT	OR	ON	2005/12/19 18:18
L49	1120	url with (token$ string)	USPAT	OR	ON	2005/12/19 18:18
L50	120	(url with (token$ string)) with (predicate query)	USPAT	OR	ON	2005/12/19 18:18
L51	28	((url with (token$ string)) with (predicate query)) and (document with retriev$)	USPAT	OR	ON	2005/12/19 18:18
L52	11	("5442784" | "5694594" | "5848407" | "5873081" | "5937422" | "5940821" | "5941944" | "5953718" | "5963940" | "5991756" | "6047126").PN.	USPAT	OR	ON	2005/12/19 18:18
L53	29	"6112203".URPN.	USPAT	OR	ON	2005/12/19 18:18
L54	29	"6112203".URPN.	USPAT	OR	ON	2005/12/19 18:18
L55	2962	(web adj page) with (relat$ retriev$)	USPAT	OR	ON	2005/12/19 18:18
L56	712	((web adj page) with (relat$ retriev$)) and (scor$5 relevanc$5 rank$4)	USPAT	OR	ON	2005/12/19 18:18

L57	100	((web adj page) with (relat$ retriev$)) same (scor$5 relevanc$5 rank$4)	USPAT	OR	ON	2005/12/19 18:18
L58	84	(((web adj page) with (relat$ retriev$)) same (scor$5 relevanc$5 rank$4)) and (content)	USPAT	OR	ON	2005/12/19 18:18
L59	78	((((web adj page) with (relat$ retriev$)) same (scor$5 relevanc$5 rank$4)) and (content)) and link$	USPAT	OR	ON	2005/12/19 18:18
L60	78	((((web adj page) with (relat$ retriev$)) same (scor$5 relevanc$5 rank$4)) and (content)) and link$5	USPAT	OR	ON	2005/12/19 18:18
L61	7	(((((web adj page) with (relat$ retriev$)) same (scor$5 relevanc$5 rank$4)) and (content)) and link$) and (url with (token$ string))	USPAT	OR	ON	2005/12/19 18:18
L62	5	"6418433".URPN.	USPAT	OR	ON	2005/12/19 18:18
L63	8	("5369577" | "5530852" | "5708829" | "5717912" | "5784608" | "5787417" | "5796952" | "5832494").PN.	USPAT	OR	ON	2005/12/19 18:18
L64	191	relevance adj feedback	USPAT	OR	ON	2005/12/19 18:18
L65	280	relevance with feedback	USPAT	OR	ON	2005/12/19 18:18
L66	6	(relevance with feedback) and ((url hyperlink) with (token$ string))	USPAT	OR	ON	2005/12/19 18:18
L67	6	(relevance adj feedback) and ((url hyperlink) with (token$ string))	USPAT	OR	ON	2005/12/19 18:18

Terms used web crawler statistical information

Found **51,107** of **167,655**

Sort results by	relevance	Save results to a Binder Search Tips □ Open results in a new window
Display results	expanded form	

Try an Advanced Search
Try this search in The ACM Guide

Results 1 - 20 of 200

Best 200 shown

Result page: **1** 2 3 4 5 6 7 8 9 10 next

Relevance scale

1 Evaluating topic-driven web crawlers
Filippo Menczer, Gautam Pant, Padmini Srinivasan, Miguel E. Ruiz
September 2001 **Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval**
Publisher: ACM Press
Full text available: pdf(210.09 KB) Additional Information: full citation, abstract, references, citings, index terms

Due to limited bandwidth, storage, and computational resources, and to the dynamic nature of the Web, search engines cannot index every Web page, and even the covered portion of the Web cannot be monitored continuously for changes. Therefore it is essential to develop effective crawling strategies to prioritize the pages to be indexed. The issue is even more important for topic-specific search engines, where crawlers must make additional decisions based on the relevance of visited pages. ...

Keywords: InfoSpiders, PageRank, Web information retrieval, best-first search, focused crawlers, performance metrics, topic driven crawling

2 On the design of a learning crawler for topical resource discovery
Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu
July 2001 **ACM Transactions on Information Systems (TOIS)**, Volume 19 Issue 3
Publisher: ACM Press
Full text available: pdf(324.39 KB) Additional Information: full citation, abstract, references, index terms

In recent years, the World Wide Web has shown enormous growth in size. Vast repositories of information are available on practically every possible topic. In such cases, it is valuable to perform topical resource discovery effectively. Consequently, several new ideas have been proposed in recent years; among them a key technique is focused crawling which is able to crawl particular topical portions of the World Wide Web quickly, without having to explore all web pages. In this paper, we propose ...

Keywords: Crawling, World Wide Web

3 Effective page refresh policies for Web crawlers
Junghoo Cho, Hector Garcia-Molina
December 2003 **ACM Transactions on Database Systems (TODS)**, Volume 28 Issue 4
Publisher: ACM Press
Full text available: pdf(345.52 KB) Additional Information: full citation, abstract, references, index terms

In this article, we study how we can maintain local copies of remote data sources "fresh," when the source data is updated autonomously and independently. In particular, we study the problem of *Web crawlers* that maintain local copies of remote Web pages for Web search engines. In this context, remote data sources (Websites) do not notify the copies (Web crawlers) of new changes, so we need to periodically *poll* the sources to maintain the copies up-to-date. Since polling the sources ...

Keywords: Web crawlers, page refresh, web search engines, world-wide web

4 Intelligent crawling on the World Wide Web with arbitrary predicates
Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu
April 2001 **Proceedings of the 10th international conference on World Wide Web**
Publisher: ACM Press
Full text available: pdf(272.60 KB) Additional Information: full citation, references, citings, index terms

Keywords: World Wide Web, crawling, querying

5 Characterizing a national community web
Daniel Gomes, Mário J. Silva
August 2005 **ACM Transactions on Internet Technology (TOIT)**, Volume 5 Issue 3
Publisher: ACM Press
Full text available: pdf(364.77 KB) Additional Information: full citation, abstract, references, index terms

This article presents a characterization of the community Web of the people of Portugal. We defined criteria for delimiting this Web based on our past experience of crawling pages related to Portugal and collected over 3.2 million documents from 46,000 sites satisfying those criteria. Our characterization was derived from this crawl. We describe the rules that we established for defining the boundaries of this community Web and the methodology used to gather statistics. Statistics cover the numb ...

Keywords: Portuguese Web, Web characterization, Web communities, Web measurements

6 Poster papers: Collaborative crawling: mining user experiences for topical resource discovery
Charu C. Aggarwal
July 2002 **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**
Publisher: ACM Press
Full text available: pdf(691.02 KB) Additional Information: full citation, abstract, references, index terms

The rapid growth of the world wide web had made the problem of topic specific resource discovery an important one in recent years. In this problem, it is desired to find web pages which satisfy a predicate specified by the user. Such a predicate could be a keyword query, a topical query, or some arbitrary contraint. Several techniques such as focussed crawling and intelligent crawling have recently been proposed for topic specific resource discovery. All these crawlers are *linkage based*, ...

7 Tools & techniques track: searching and IR: Downloading textual hidden web content through keyword queries
Alexandros Ntoulas, Petros Zerfos, Junghoo Cho
June 2005 **Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries**
Publisher: ACM Press
Full text available: pdf(278.40 KB) Additional Information: full citation, abstract, references, index terms

An ever-increasing amount of information on the Web today is available only through search interfaces: the users have to type in a set of keywords in a search form in order to access the pages from certain Web sites. These pages are often referred to as the *Hidden Web* or the *Deep Web*. Since there are no static links to the Hidden Web pages, search engines cannot discover and index such pages and thus do not return them in the results. However, according to recent studies, the conte ...

Keywords: adaptive algorithm, deep web crawler, hidden web crawling, keyword queries, query selection

8 A language and character set determination method based on N-gram statistics
Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, Yoshihide Chubachi
September 2002 **ACM Transactions on Asian Language Information Processing (TALIP)**, Volume 1 Issue 3
Publisher: ACM Press
Full text available: pdf(94.47 KB) Additional Information: full citation, abstract, references, index terms

An N-gram-based language, script, and encoding scheme-detection method is introduced in this article. The method detects language, script, and encoding schemes using a target text document encoded by computer by checking how many byte sequences of the target match the byte sequences that can appear in the texts belonging to a language, script, and encoding scheme. This detection mechanism is different from conventional N-gram-based methods in that its threshold for any category is uniquely prede ...

Keywords: N-gram, Unicode, character set, corpus-based analysis, local language site, natural languages, text categorization

9 Information retrieval on the web
Mei Kobayashi, Koichi Takeda
June 2000 **ACM Computing Surveys (CSUR)**, Volume 32 Issue 2
Publisher: ACM Press
Full text available: pdf(213.89 KB) Additional Information: full citation, abstract, references, citings, index terms

In this paper we review studies of the growth of the Internet and technologies that are useful for information search and retrieval on the Web. We present data on the Internet from several different sources, e.g., current as well as projected number of users, hosts, and Web sites. Although numerical figures vary, overall trends cited by the sources are consistent and point to exponential growth in the past and in the coming decade. Hence it is not surprising that about 85% of Internet user ...

Keywords: Internet, World Wide Web, clustering, indexing, information retrieval, knowledge management, search engine

10 Estimating frequency of change
Junghoo Cho, Hector Garcia-Molina
August 2003 **ACM Transactions on Internet Technology (TOIT)**, Volume 3 Issue 3
Publisher: ACM Press
Full text available: pdf(353.56 KB) Additional Information: full citation, abstract, references, citings, index terms

Many online data sources are updated autonomously and independently. In this article, we make the case for estimating the change frequency of data to improve Web crawlers, Web caches and to help data mining. We first identify various scenarios, where different applications have different requirements on the accuracy of the estimated frequency. Then we develop several "frequency estimators" for the identified scenarios, showing analytically and experimentally how precise they are. In many cases, ...

Keywords: Change frequency estimation, Poisson process

11 Learning probabilistic models of the Web (poster session)
Thomas Hofmann
July 2000 **Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval**
Publisher: ACM Press
Full text available: pdf(378.00 KB) Additional Information: full citation, abstract, references, citings, index terms

In the World Wide Web, myriads of hyperlinks connect documents and pages to create an unprecedented, highly complex graph structure - the Web graph. This paper presents a novel approach to learning probabilistic models of the Web, which can be used to make reliable predictions about connectivity and information content of Web documents. The proposed method is a probabilistic dimension reduction technique which recasts and unites Latent Semantic Analysis and Kleinberg's Hubs-and-Authorities al ...

12 Searching the Web
August 2001 **ACM Transactions on Internet Technology (TOIT)**, Volume 1 Issue 1
Publisher: ACM Press
Full text available: pdf(319.98 KB) Additional Information: full citation, abstract, references, citings, index terms, review

We offer an overview of current Web search engine design. After introducing a generic search engine architecture, we examine each engine component in turn. We cover crawling, local Web page storage, indexing, and the use of link analysis for boosting search performance. The most common design and implementation techniques for each of these components are presented. For this presentation we draw from the literature and from our own experimental search engine testbed. Emphasis is on introduci ...

Keywords: HITS, PageRank, authorities, crawling, indexing, information retrieval, link analysis, search engine

13 Building a distributed full-text index for the web
July 2001 **ACM Transactions on Information Systems (TOIS)**, Volume 19 Issue 3
Publisher: ACM Press
Full text available: pdf(651.72 KB) Additional Information: full citation, abstract, references, index terms, review

We identify crucial design issues in building a distributed inverted index for a large collection of Web pages. We introduce a novel pipelining technique for structuring the core index-building system that substantially reduces the index construction time. We also propose a storage scheme for creating and managing inverted files using an embedded database system. We suggest and compare different strategies for collecting global statistics from distributed inverted indexes. Finally, we present pe ...

Keywords: Distributed indexing, Embedded databases, Inverted files, Pipelining, Text retrieval

14 Web crawling and exploration: Probabilistic models for focused web crawling
Hongyu Liu, Evangelos Milios, Jeannette Janssen
November 2004 **Proceedings of the 6th annual ACM international workshop on Web information and data management**
Publisher: ACM Press
Full text available: pdf(384.56 KB) Additional Information: full citation, abstract, references, index terms

A Focused crawler must use information gleaned from previously crawled page sequences to estimate the relevance of a newly seen URL. Therefore, good performance depends on powerful modelling of context as well as the current observations. Probabilistic models, such as Hidden Markov Models(HMMs) and Conditional Random Fields(CRFs), can potentially capture both formatting and context. In this paper, we present the use of HMM for focused web crawling, and compare it with Best-First strategy. Fur ...

Keywords: conditional random fields, focused crawling, hidden Markov models, web graph, world wide web

15 An adaptive model for optimizing performance of an incremental web crawler
Jenny Edwards, Kevin McCurley, John Tomlin
April 2001 **Proceedings of the 10th international conference on World Wide Web**
Publisher: ACM Press

Full text available: pdf(201.50 KB) Additional Information: full citation, references, citings, index terms

Keywords: crawler, incremental crawler, optimization, scalability

16 Short papers: Discovery of ads web hosts through traffic data analysis

V. Bacarella, F. Giannotti, M. Nanni, D. Pedreschi

June 2004 **Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery**

Publisher: ACM Press

Full text available: pdf(189.30 KB) Additional Information: full citation, abstract, references

One of the most actual problems on web crawling -- the most expensive task of any search engine, in terms of time and bandwidth consumption -- is the detection of useless segments of Internet. In some cases such segments are purposely created to deceive the crawling engine while, in others, they simply do not contain any useful information. Currently, the typical approach to the problem consists in using a human-compiled *blacklist* of sites to avoid (e.g., advertising sites and web counter ...

17 Learning to crawl: Comparing classification schemes

Gautam Pant, Padmini Srinivasan

October 2005 **ACM Transactions on Information Systems (TOIS)**, Volume 23 Issue 4

Publisher: ACM Press

Full text available: pdf(940.75 KB) Additional Information: full citation, abstract, references, index terms

Topical crawling is a young and creative area of research that holds the promise of benefiting from several sophisticated data mining techniques. The use of classification algorithms to guide topical crawlers has been sporadically suggested in the literature. No systematic study, however, has been done on their relative merits. Using the lessons learned from our previous crawler evaluation studies, we experiment with multiple versions of different classification schemes. The crawling process is ...

Keywords: Topical crawlers, classifiers, focused crawlers, machine learning

18 Data integrity: Web application security assessment by fault injection and behavior monitoring

Yao-Wen Huang, Shih-Kun Huang, Tsung-Po Lin, Chung-Hung Tsai

May 2003 **Proceedings of the 12th international conference on World Wide Web**

Publisher: ACM Press

Full text available: pdf(4.53 MB) Additional Information: full citation, abstract, references, citings, index terms

As a large and complex application platform, the World Wide Web is capable of delivering a broad range of sophisticated applications. However, many Web applications go through rapid development phases with extremely short turnaround time, making it difficult to eliminate vulnerabilities. Here we analyze the design of Web application security assessment mechanisms in order to identify poor coding practices that render Web applications vulnerable to attacks such as SQL injection and cross-site scr ...

Keywords: black-box testing, complete crawling, fault injection, security assessment, web application testing

19 Web search 1: Topic-oriented collaborative crawling

Chiasen Chung, Charles L. A. Clarke

November 2002 **Proceedings of the eleventh international conference on Information and knowledge management**

Publisher: ACM Press

Full text available: pdf(179.28 KB) Additional Information: full citation, abstract, references, index terms

A major concern in the implementation of a distributed Web crawler is the choice of a strategy for partitioning the Web among the nodes in the system. Our goal in selecting this strategy is to minimize the overlap between the activities of individual nodes. We propose a topic-oriented approach, in which the Web is partitioned into general subject areas with a crawler assigned to each. We examine design alternatives for a topic-oriented distributed crawler, including the creation of a Web page cl ...

Keywords: distributed systems, text categorization, web crawling

20 Organizing topic-specific web information
Sougata Mukherjea
May 2000 **Proceedings of the eleventh ACM on Hypertext and hypermedia**
Publisher: ACM Press
Full text available: pdf(183.02 KB) Additional Information: full citation, references, citings, index terms

Keywords: World-Wide Web, abstraction hierarchy, graph algorithms, information visualization, topic management

Results 1 - 20 of 200 Result page: **1** 2 3 4 5 6 7 8 9 10 next